

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

Translated Excerpt of Citation 3

Japanese Patent Laid-Open Publication No. HEI 5-219932

* * * * *

Paragraphs [0015] – [0021]:

“[0015]

[Action]

According to the genetic information examining device 1 of the present invention shown in Fig. 1(a), when character sequence expression of the amino acid arrangement for examination is "ABCB DAB" while character sequence expression of the amino acid arrangement for comparison is "BDCABA", for example, detecting means 10 detects that the number of the longest common characters in these two character sequences is "4", and calculating means 11, in response to the detection result, calculates the rate value of 57% ($=4/7$), when based on the character sequence length of the amino acid arrangement for examination, and when based on the character sequence length of the amino acid arrangement for comparison, the rate value of 67% ($=4/6$) is calculated.

[0016]

The output controlling means 12 outputs the calculated rate value to an outputting device 2 to notify a user of the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison. On the other hand, evaluating means 13 compares the calculated rate value with the regular standard to evaluate the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison mechanically, and to perform the above-mentioned homology examination.

[0017]

As described above, the genetic information examining device 1 of the present invention according to Fig. 1 (a) has the composition which calculates the number of the longest common characters of the amino acid arrangement for examination expressed in a character sequence and the amino acid arrangement for comparison expressed in a character sequence. Since it is the composition of evaluating the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison according to this number of the longest common characters, compared with

a comparison method of the amino acid arrangement by dynamic programming technique, the similarity of two amino acid arrangements can be evaluated with small memory capacity and at high speed.

[0018]

According to the genetic information examining device 1 of the present invention shown in Fig. 1(b), when character sequence expression of the amino acid arrangement for examination is "ABCB DAB" while character sequence expression of the amino acid arrangement for comparison is "BDCABA", detecting means 20 detects that the longest common portion sequences in these two character sequences are "BDAB", "BCBA", "BDAB" and "BCAB". In response to the detection result, specifying means 21 specifies the arrangement position of each longest common portion sequence in the amino acid arrangement for examination and the amino acid arrangement for comparison, and also specifies the character sequence length existing between the arrangement positions.

[0019]

Then, the output controlling means 22 outputs the longest common portion sequence detected by the detecting means 20 as it is, outputs the character sequence length specified by the specifying means 21 so that it is corresponding to the longest common portion sequence, and moreover, outputs these two amino acid arrangements while aligning so that the longest common portion sequences of the amino acid arrangement for examination and the amino acid arrangement for comparison are corresponding to each other, according to the arrangement position specified by the specifying means 21. In such a manner, the output controlling means 22 notifies the user of the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison.

[0020]

On the other hand, when the evaluating means 23 evaluates whether the amino acid arrangement for comparison is included in the amino acid arrangement for examination in case the amino acid arrangement for comparison is expressed in the successive character sequence or in character sequence containing non-specified character sequence between arrangement positions, the evaluating means 23 evaluates mechanically whether the longest common portion sequence detected by detecting means 20 and the amino acid arrangement for comparison coincide or not, taking into consideration the

specified result by the specifying means 21, and performs the above-described motif examination.

[0021]

As described above, the genetic information examining device 1 of the present invention according to Fig. 1 (b) has the composition specifying the longest common portion sequence of the amino acid arrangement for examination expressed in a character sequence and the amino acid arrangement for comparison expressed in a character sequence. Since it is the composition of evaluating the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison according to the longest common portion sequence, compared with a comparison method of the amino acid arrangement by dynamic programming technique, the similarity of two amino acid arrangement can be evaluated with small memory capacity and at high speed. ”

* * * * *

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平5-219932

(43)公開日 平成5年(1993)8月31日

(51)Int.Cl.⁵

識別記号

庁内整理番号

FI

技術表示箇所

C 1 2 M 1/00

A 9050-4B

G 0 6 F 15/40

5 1 0 E 7060-5L

15/42

A 7060-5L

審査請求 未請求 請求項の数5(全17頁)

(21)出願番号

特願平4-21012

(22)出願日

平成4年(1992)2月6日

(71)出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中1015番地

(72)発明者 大矢 真弓

神奈川県川崎市中原区上小田中1015番地

富士通株式会社内

(72)発明者 相川 聖一

神奈川県川崎市中原区上小田中1015番地

富士通株式会社内

(74)代理人 弁理士 森田 寛 (外1名)

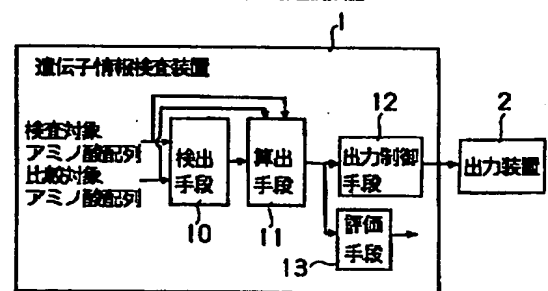
(54)【発明の名称】 遺伝子情報検査装置

(57)【要約】

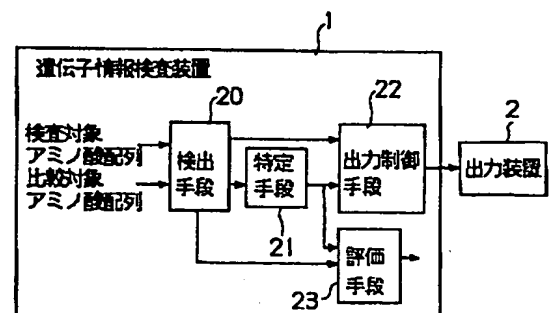
【目的】本発明は、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置に関し、簡単な処理機構に従って類似性を評価することのできるようにすることを目的とする。

【構成】アミノ酸を文字で表現する構成を採り、かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段10と、この検出手段10により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段11とを備えるように構成し、また、アミノ酸を文字で表現する構成を採り、かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段(20)を備えるように構成する。

本発明の原理構成図



(a)



(b)

【特許請求の範囲】

【請求項1】 検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置において、

アミノ酸を文字で表現する構成を採り、

かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段(10)と、

上記検出手段(10)により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段(11)とを備えることを、

特徴とする遺伝子情報検査装置。

【請求項2】 検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置において、

アミノ酸を文字で表現する構成を採り、

かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段(20)を備えることを、

特徴とする遺伝子情報検査装置。

【請求項3】 請求項2記載の遺伝子情報検査装置において、

検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列の配列位置を特定する特定手段(21)を備えることを、

特徴とする遺伝子情報検査装置。

【請求項4】 請求項3記載の遺伝子情報検査装置において、

出力装置に対して2つのアミノ酸配列を出力する構成を採って、この出力時に、特定手段(21)により特定される配列位置に従って、この2つのアミノ酸配列の持つ最長共有部分列が対応付けられるべく該アミノ酸配列をアラインメントして出力していくよう処理することを、

特徴とする遺伝子情報検査装置。

【請求項5】 請求項3記載の遺伝子情報検査装置において、

出力装置に対して2つのアミノ酸配列の持つ最長共有部分列を出力する構成を採って、この出力時に、特定手段(21)により特定される配列位置に従って規定される該配列位置間の持つ文字列長を、該最長共有文字列に対応付けて出力していくよう処理することを、

特徴とする遺伝子情報検査装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置に関し、特に、簡単な処理機構に従って類似性を評価することのできる遺伝子情報検査装置に関するものである。

【0002】医薬品の開発等に必要なタンパク質工学の分野では、分子生物学の進歩に伴って大量に遺伝情報が蓄積され始め現在データベース化が進んでいる。これらの大量に蓄積された遺伝情報から、タンパク質の構造や機能等の生物学的に意味のある情報を抽出していくことが要求されている。この抽出処理は、高速処理を可能にするためにも、できる限り簡単な処理機構により実現していくことが好ましいのである。

【0003】

【従来の技術】遺伝子の本体はDNAであり、A(アデニン)、T(チミン)、C(シトシン)、G(グアニン)という4つの塩基で構成される塩基配列として表現される。また、生体を構成するアミノ酸は約20種類あり、これまでに、塩基配列中の3つの塩基の並びと各アミノ酸とが対応することが解明されている。従って、生体内では、DNAの塩基配列に従ってアミノ酸が合成され、合成されたアミノ酸が折り畳まれることによってタンパク質が形作られることになる。

【0004】上述したように、分子生物学の発展に伴い、塩基やアミノ酸の配列の決定法が確立したことによって、塩基配列データ、アミノ酸配列データ等の遺伝情報が大量に蓄積され始めている。このため、遺伝子情報処理の分野では、蓄積された膨大な遺伝情報の中から、タンパク質の構造や機能等に関する生物情報をいかにして抽出するかが中心課題となってきた。

【0005】このような生物情報を抽出する際の基本的手法は、アミノ酸の配列を比較することである。これは、アミノ酸の配列が類似していることは、生物学的機能にも類似があると考えられているためである。

【0006】このようなことを背景にして、評価対象のアミノ酸配列の機能を推定するために、機能が解明されている既知のアミノ酸配列データベースから、評価対象のアミノ酸配列と類似するアミノ酸配列を検索する相同性探索や、比較するアミノ酸配列間での違いと類似性が明確になるようにアミノ酸配列を並び変えるアラインメントが行われるようになってきている。

【0007】また、アミノ酸配列の中で生物にとって重要な機能をコードしている領域は、進化の過程でも保存されていると考えられている。例えば、異なる生物種で同じ機能を持つタンパク質のアミノ酸配列を比較すると、共通に存在する配列パターンがあることが知られている。このような配列パターンはモチーフと呼ばれている。これから、アミノ酸配列中にどのようなモチーフが含まれているかを調べることによって、タンパク質の性質や機能を解明することができるだけでなく、既存のタンパク質に対する強化、機能の付加、新しいタンパク質の合成等、多岐に渡ってタンパク質工学の分野に応用することができる。これから、モチーフを検索することが行われるようになってきている。

【0008】従来、2つのアミノ酸配列を比較する方法

としては、音声認識処理等で用いられているダイナミックプログラミング手法が用いられている。

【0009】

【発明が解決しようとする課題】しかしながら、ダイナミックプログラミング手法によるアミノ酸配列の比較方法では、2次元的にアミノ酸配列を比較していくために、大きなメモリ容量が必要になるとともに、処理時間も多くなるという問題点があった。

【0010】本発明はかかる事情に鑑みてなされたものであって、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を簡単な処理機構に従って評価することのできる新たな遺伝子情報検査装置の提供を目的とするものである。

【0011】

【課題を解決するための手段】図1(a)に本発明の第1の発明の原理構成、図1(b)に本発明の第2の発明の原理構成を図示する。

【0012】図1(a)(b)中、1は本発明を具備する遺伝子情報検査装置であって、アミノ酸を文字で表現する構成を採って、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価するもの、2は遺伝子情報検査装置1に接続される出力装置である。

【0013】図1(a)に従う本発明の遺伝子情報検査装置1は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段10と、検出手段10により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段11と、算出手段11の算出する割合値を出力装置2に出力する出力制御手段12と、算出手段11の算出する割合値に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する評価手段13とを備える。

【0014】図1(b)に従う本発明の遺伝子情報検査装置1は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段20と、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列の配列位置を特定するとともに、その特定結果に従って、その配列位置間に存在する文字列長を特定する特定手段21と、検出手段20の検出結果や特定手段21の特定結果を出力装置2に出力する出力制御手段22と、検出手段20の検出結果や特定手段21の特定結果に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する評価手段23とを備える。

【0015】

【作用】図1(a)に従う本発明の遺伝子情報検査装置1では、例えば、検査対象アミノ酸配列の文字列表現が“ABCBDAB”で、比較対象アミノ酸配列の文字列表現が“BDCABA”である場合に、検出手段10

は、この2つの文字列での最長共有文字数が“4”であることを検出し、算出手段11は、この検出結果を受けて、検査対象アミノ酸配列の文字列長を基準にする場合には57%($=4 \div 7$)という割合値を算出し、比較対象アミノ酸配列の文字列長を基準にする場合には67%($=4 \div 6$)という割合値を算出する。

【0016】そして、出力制御手段12は、この算出された割合値を出力装置2に出力していくことで、ユーザに対して、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性の評価値を通知し、一方、評価手段13は、この算出された割合値を規定の基準値と比較することで、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を機械的に評価して上述の相同性検査を実行していく。

【0017】このように、図1(a)に従う本発明の遺伝子情報検査装置1は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を算出する構成を採って、この最長共有文字数に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミング手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に2つのアミノ酸配列の類似性を評価することができるのである。

【0018】図1(b)に従う本発明の遺伝子情報検査装置1では、例えば、検査対象アミノ酸配列の文字列表現が“ABCBDAB”で、比較対象アミノ酸配列の文字列表現が“BDCABA”である場合に、検出手段20は、この2つの文字列での最長共有部分列が“BDA B”、“BCBA”、“BDAB”、“BCAB”であることを検出する。この検出結果を受けて、特定手段21は、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ各最長共有部分列の配列位置を特定するとともに、その配列位置間に存在する文字列長を特定する。

【0019】そして、出力制御手段22は、検出手段20の検出する最長共有部分列をそのまま出力したり、特定手段21により特定される文字列長をこの最長共有部分列に対応付けて出力したり、特定手段21により特定される配列位置に従って、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列が対応付けられるべくこの2つのアミノ酸配列をアラインメントして出力したりしていくことで、ユーザに対して、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を通知する。

【0020】一方、評価手段23は、比較対象アミノ酸配列が連続する文字列で表現される場合や、規定されない文字列を配列位置間に含む文字列のもので表現される場合にあって、検査対象アミノ酸配列にこの比較対象アミノ酸配列が含まれているか否かを評価していくときには、特定手段21の特定結果を考慮しつつ、検出手段2



0の検出する最長共有部分列と比較対象アミノ酸配列とが一致するか否かを機械的に評価していくことで上述のモチーフ検査を実行していく。

【0021】このように、図1(b)に従う本発明の遺伝子情報検査装置1は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を特定する構成を採って、この最長共有部分列に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミング手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に2つのアミノ酸配列の類似性を評価することができるのである。

【0022】

【実施例】以下、実施例に従って本発明を詳細に説明する。図2に、本発明を実装する遺伝子情報検査装置1の一実施例を図示する。図中、40は遺伝子情報検査装置1に接続される入力装置、41は入力装置40の備えるキーボードやマウス等の対話装置、42は遺伝子情報検査装置1に接続されるディスプレイ装置、50は文字列で表現されるアミノ酸配列情報を管理するアミノ酸配列データベース、60は文字列で表現されるモチーフ配列情報を管理するモチーフデータベースである。

【0023】この実施例の遺伝子情報検査装置1は、入力装置40から入力されてくるアミノ酸配列の文字列と、アミノ酸配列データベース50やモチーフデータベース60から与えられるアミノ酸配列の文字列との間の最長共有文字数や、最長共有部分列(LCS: longest common subsequence)や、最長共有部分列の展開位置を検出するLCS検出部30と、LCS検出部30の結果に従って、LCS検出部30の検出対象となった2つのアミノ酸配列の相同性を判定する相同性判定部31と、相同性判定部31の検出結果に従って、入力装置40から入力されてくるアミノ酸配列と相同なアミノ酸配列をアミノ酸配列データベース50から検索する相同性探索部32と、LCS検出部30の結果に従って、入力装置40から入力されてくるアミノ酸配列と相同なモチーフ配列をモチーフデータベース60から検索するモチーフ探索部33と、LCS検出部30の検出結果に従って、入力装置40から入力されてくるアミノ酸配列の文字列と、アミノ酸配列データベース50やモチーフデータベース60から与えられるアミノ酸配列の文字列とをアラインメントするアラインメント部34と、各処理部の処理結果をディスプレイ装置42に表示する表示部35とを備える。

【0024】次に、図3ないし図5に示す処理フローに従って、LCS検出部30の実行する処理について詳細に説明する。ここで、図3に示す処理フローは、検査対象となる2つのアミノ酸配列の持つ最長共有文字数を検出するための処理フローであり、図4及び図5に示す処

理フローは、検査対象となる2つのアミノ酸配列の持つ最長共有部分列と、その展開位置を検出するための処理フローである。

【0025】LCS検出部30は、文字列1で表現されるアミノ酸配列と、文字列2で表現されるアミノ酸配列との間の最長共有文字数を検出する場合には、図3の処理フローに示すように、先ず最初に、ステップ1で、文字列1から1文字ずつ読み込んで、文字列1中での各文字の出現位置を示す表を作成する。

【0026】この出現表は、例えば、A～Zまでのアルファベットに対応した配列に、各文字の出現位置をポインタで連結することによって実現するものであって、例えば、文字列1のアミノ酸配列が“ABCBDAB”で表現される場合には、図6に示すように、“A”が6番目と1番目に出現し、“B”が7番目と4番目と2番目に出現し、“C”が3番目に出現し、“D”が5番目に出現するというように作成する。そして、このステップ1では、更に、以下の処理で用いる文字列1と同じサイズを持つ配列S[i]の初期化処理を実行して、各エントリにゼロ値を設定する。

【0027】次に、ステップ2で、文字列2から1文字を読み込み、ステップ1で作成した出現表を参照して、その文字の文字列1での出現位置rを特定する。続いて、ステップ3で、用意されている配列S[i]のr番目のS[r]のエントリデータと、(r-1)番目のS[r-1]のエントリデータとが等しいか否かを判断する。

【0028】このステップ3で、S[r]とS[r-1]とが等しいと判断するときには、ステップ4に進んで、r番目以上で、かつS[r]のエントリデータと等しい値のエントリデータを持つ配列S[i]に“1”を加算し、続くステップ5で、文字列2の最後の文字までの処理を終了したのか否かを判断して、終了していないことを判断するときには、ステップ2に戻っていく。一方、ステップ3で、S[r]とS[r-1]とが等しくないと判断するときには、ステップ4の加算処理を実行することなく、直ちにステップ5に進んでいく。

【0029】ここで、ステップ2で読み込んだ文字列2の文字が文字列1で複数回出現する場合には、出現位置rの大きい順にステップ3及びステップ4の処理を実行していくことになる。

【0030】そして、ステップ5で、文字列2の最後の文字までの処理を終了したことを判断すると、ステップ6に進んで、配列S[i]の最終要素のS[m]のエントリデータKmaxを最長共有文字数として出力していく。

【0031】この処理フローの実行により、例えば、文字列1のアミノ酸配列が“ABCBDAB”で表現され、文字列2のアミノ酸配列が“BDCABA”で表現される場合には、文字列2の第1番目の文字Bの読込処理に従って、図6の出現表から“r=7, 4, 2”が特定されて、図7(a)に示すように配列S[i]のエント

リデータが更新され、文字列2の第2番目の文字Dの読込処理に従って、図6の出現表から“ $r=5$ ”が特定されて、図7(b)に示すように配列 $S[i]$ のエントリデータが更新され、文字列2の第3番目の文字Cの読込処理に従って、“ $r=3$ ”が特定されて、図8(a)に示すように配列 $S[i]$ のエントリデータが更新され、文字列2の第4番目の文字Aの読込処理に従って、図6の出現表から“ $r=6, 1$ ”が特定されて、図8(b)に示すように配列 $S[i]$ のエントリデータが更新される。

【0032】そして、文字列2の第5番目の文字Bの読込処理に従って、図6の出現表から“ $r=7, 4, 2$ ”が特定されて、図9(a)に示すように配列 $S[i]$ のエントリデータが更新され、文字列2の第6番目の文字Aの読込処理に従って、図6の出現表から“ $r=6, 1$ ”が特定されて、図9(b)に示すように配列 $S[i]$ のエントリデータが更新されていって、最終的に、“4”という最長共有文字数が特定されることになる。なお、図7ないし図9に示す配列 $S[i]$ では、システムの便宜上、文字列1より1文字多いサイズを持つ配列 $S[i]$ に従うもので示してある。

【0033】次に、図4及び図5に従って、検査対象となる2つのアミノ酸配列の持つ最長共有部分列と、その展開位置を検出するための処理について説明する。LCS検出部30は、文字列1で表現されるアミノ酸配列と、文字列2で表現されるアミノ酸配列との間の最長共有部分列と、その展開位置とを検出する場合には、図4の処理フローに示すように、先ず最初に、ステップ10で、文字列1から1文字ずつ読み込んで、文字列1中での各文字の出現位置を示す表を作成する。すなわち、図6で説明した出現表を作成するのである。そして、このステップ1では、更に、以下の処理で用いる文字列1と同じサイズを持つ配列 $S[i]$ の初期化処理を実行して、各エントリにゼロ値を設定するとともに、以下の処理で用いる最長共有文字数と同じサイズを持つ配列 $data[k]$ の初期化処理を実行して、各エントリが何もポイントしないように設定する。

【0034】次に、ステップ11で、文字列2から1文字(j 番目の文字)を読み込み、ステップ10で作成した出現表を参照して、その文字の文字列1での出現位置 r を特定する。続いて、ステップ12で、用意されている配列 $S[i]$ の r 番目の $S[r]$ のエントリデータと、($r-1$)番目の $S[r-1]$ のエントリデータとが等しいか否かを判断する。このステップ12で、 $S[r]$ と $S[r-1]$ とが等しいと判断するときには、ステップ13に進んで、 r 番目以上で、かつ $S[r]$ のエントリデータと等しい値のエントリデータを持つ配列 $S[i]$ に“1”を加算し、一方、等しくないと判断するときには、図5の処理フローのステップ17の処理に進んで、この加算処理を実行しないよう処理する。ここで、ステップ11で読み込んだ文字列2の文字が文字列1で複数回出現する場合に

は、出現位置 r の大きい順にステップ12及びステップ13の処理を実行していくことになる。

【0035】このようにして設定される $S[r]$ のエントリデータの値 k が、文字列1の r 番目の文字までの文字列と、文字列2の j 番目の文字までの文字列との間の最長共有文字数となる。このように、LCS検出部30は、最長共有部分列を検出していく場合にも、図3の処理フローで説明した最長共有文字数を検出していく処理を実行していくものである。

【0036】ステップ13の処理を実行すると、続いて、ステップ14で、得られた $S[r]$ のエントリデータである最長共有文字数 k に従って、文字列1での展開位置 r と、文字列2での展開位置 j との対データ(r, j)を配列 $data[k]$ に格納する。ここで、配列 $S[i]$ が前回の処理サイクルのものから変化していないときには、この格納処理を実行しないように処理する。最長共有部分列は、以下の処理に従って、このデータ構造を連結していくことで求められることになる。

【0037】続いて、図5の処理フローに移って、ステップ15で、 $data[k-1]$ に格納された文字位置 r', j' を参照して、

$$r' < r, \quad j' < j$$

が成立するか否かを判断し、成立すると判断するときには、文字位置の逆転が起こらないことに対応して、ステップ16に進んで、文字位置 r', j' を次候補となるものとしてポインタを張って登録する。そして、続くステップ17で、文字列2の最後の文字までの処理を終了したのか否かを判断して、終了していないことを判断するときには、図4の処理フローのステップ11に戻っていく。一方、ステップ15で、上述の関係式が成立しないと判断するときには、ステップ16の処理を実行することなく、直ちにステップ17の処理に入っていく。

【0038】そして、ステップ17で、文字列2の最後の文字までの処理を終了したことを判断すると処理を終了する。この図4及び図5の処理フローの実行により、上述のように、文字列1のアミノ酸配列が“ABCBDAB”で表現され、文字列2のアミノ酸配列が“BDCABA”で表現される場合には、文字列2の第1番目($j=1$)の文字Bの読込処理に従って、図6の出現表から“ $r=7, 4, 2$ ”が特定されて、図7(a)に示したように、“ $r=7$ ”に従って“ $S[7]=1$ ”が特定されることで $data[1]$ に(7, 1)が格納され、“ $r=4$ ”に従って“ $S[4]=1$ ”が特定されることで $data[1]$ に(4, 1)が格納され、“ $r=2$ ”に従って“ $S[2]=1$ ”が特定されることで $data[1]$ に(2, 1)が格納される。

【0039】そして、文字列2の第2番目($j=2$)の文字Dの読込処理に従って、図6の出現表から“ $r=5$ ”が特定されて、図7(b)に示したように、“ $S[5]=2$ ”が特定されることで $data[2]$ に(5, 2)が

格納される。そして、文字列2の第3番目 ($j=3$) の文字Cの読込処理に従って、図6の出現表から“ $r=3$ ”が特定されて、図8(a)に示したように、“ $S[3]=2$ ”が特定されることでdata[2]に(3, 3)が格納される。そして、文字列2の第4番目 ($j=4$) の文字Aの読込処理に従って、図6の出現表から“ $r=6, 1$ ”が特定されて、図8(b)に示したように、“ $r=6$ ”に従って“ $S[6]=3$ ”が特定されることでdata[3]に(6, 4)が格納され、“ $r=1$ ”に従って“ $S[1]=1$ ”が特定されることでdata[1]に(1, 4)が格納される。

【0040】そして、文字列2の第5番目 ($j=5$) の文字Bの読込処理に従って、図6の出現表から“ $r=7, 4, 2$ ”が特定されて、図9(a)に示したように、“ $r=7$ ”に従って“ $S[7]=4$ ”が特定されることでdata[4]に(7, 5)が格納され、“ $r=4$ ”に従って“ $S[4]=3$ ”が特定されることでdata[3]に(4, 5)が格納され、“ $r=2$ ”に従って“ $S[2]=2$ ”が特定されることでdata[2]に(2, 5)が格納される。そして、文字列2の第6番目 ($j=6$) の文字Aの読込処理に従って、図6の出現表から“ $r=6, 1$ ”が特定されて、図9(b)に示したように、“ $r=6$ ”に従って“ $S[6]=4$ ”が特定されることでdata[4]に(6, 6)が格納される。なお、図9(b)から分かるように、“ $r=6$ ”と“ $r=1$ ”とで配列S[i]に変化がないことから、(1, 6)の格納処理は実行されない。

【0041】そして、これらの文字位置情報(文字列1と文字列2の持つ同一文字の展開位置を表示する)は、data[k-1]に格納されたものと、data[k]に格納されたものとで文字位置の逆転が起こらない場合には、それらの間でポインタが張られていくことで、図10のように、data[k]に格納されるのである。

【0042】最長共有部分列は、このdata[k]に格納される文字位置情報のポインタを辿っていくことで特定されることになる。すなわち、図10の例で説明するならば、“data[4]の(7, 5)→data[3]の(6, 4)→data[2]の(5, 2)→data[1]の(4, 1)”という連結に従って、最長共有部分列BDABと、文字列1及び文字列2におけるその展開位置が特定され、“data[4]の(7, 5)→data[3]の(6, 4)→data[2]の(5, 2)→data[1]の(2, 1)”という連結に従って、最長共有部分列BDABと、文字列1及び文字列2におけるその展開位置が特定され、“data[4]の(7, 5)→data[3]の(6, 4)→data[2]の(3, 3)→data[1]の(2, 1)”という連結に従って、最長共有部分列BCABと、文字列1及び文字列2におけるその展開位置が特定され、“data[4]の(6, 6)→data[3]の(4, 5)→data[2]の(3, 3)→data[1]の(2, 1)”という連結に従って、最長共有部分列BC

BAと、文字列1及び文字列2におけるその展開位置が特定されるのである。

【0043】図11及び図12に、LCS検出部30が、この連結を辿っていくことで最長共有部分列を特定していくときに実行する処理フローを図示する。次に、図2に示した遺伝子情報検査装置1の各処理部が、このLCS検出部30の検出する最長共有文字数と、最長共有部分列及びその展開位置とを受けて実行することになる処理について説明する。

10 【0044】相同性判定部31は、LCS検出部30が入力装置40から入力されてくるアミノ酸配列の文字列(以下、入力アミノ酸配列と称する)と、アミノ酸配列データベース50やモチーフデータベース60から与えられるアミノ酸配列の文字列との間の最長共有文字数を検出すると、その最長共有文字数と入力アミノ酸配列の文字列長との割合値を検出して、その割合値が規定の基準値よりも大きい場合には、入力アミノ酸配列が、アミノ酸配列データベース50やモチーフデータベース60から与えられるアミノ酸配列と相同であると判定し、基準値よりも小さい場合には、相同でないと判定する。

【0045】相同性探索部32は、相同性判定部31の判定結果を利用して、入力アミノ酸配列と相同なアミノ酸配列をアミノ酸配列データベース50から検索する。そして、相同の関係にある場合には、相同性判定部31により算出された割合値と、LCS検出部30により検出された最長共有部分列とを表示部35を介してディスプレイ装置42に表示する。

30 【0046】図13に、この表示例の一例を図示する。この表示例は、ヒトチクロームcとバクテリアチクロームcという2つのアミノ酸配列の処理結果を表示するものであって、最長共有部分列については、両者のアミノ酸配列にどのような文字間隔をもって配置されているかを示す表示形態に従って表示する構成を採っている。すなわち、“GD{x3, 3}G{x0, 1}K{x0, 2}・・・”と表示する形態を採って、ヒトチクロームcでは、“GD”の後3文字については一致しない文字が続いて、その後に“G”が続いて、その後直ぐに“K”が続き、一方、バクテリアチクロームcでは、“GD”の後3文字については一致しない文字が続いて、その後に“G”が続いて、その後1文字については一致しない文字が続いて、その後に“K”が続くというように表示するものである。

40 【0047】モチーフ探索部33は、相同性判定部31の判定結果を利用して、先ず最初に、入力アミノ酸配列と相同なモチーフ配列をモチーフデータベース60から検索し、続いて、LCS検出部30の検出する最長共有部分列と、その配列位置間の持つ文字列長とに従って、この相同の関係にあるモチーフ配列が本来のモチーフ配列であるか否かを判定する。例えば、“L”の後に規定されない文字が6文字続いてその後に“L”が続き、こ

の“L”の総個数が5個となるロイシンジッパーというモチーフ配列が、規定されない6文字まで含めて入力アミノ酸配列に含まれているか否かをLCS検出部30の検出結果に従ってチェックしていくのである。そして、モチーフ探索部33は、入力アミノ酸配列がモチーフ配列を持つ場合には、入力アミノ酸配列とモチーフ配列とを表示部35を介してディスプレイ装置42に表示する。図14に、ロイシンジッパーを持つラット卵細胞カリウムチャンネルの表示例を図示する。

【0048】アラインメント部34は、LCS検出部30の検出する最長共有部分列と、その展開位置とを受けて、入力アミノ酸配列と、アミノ酸配列データベース50やモチーフデータベース60から与えられるアミノ酸配列の持つ最長共有部分列が対応付けられるべく、この2つのアミノ酸配列をアラインメントして表示部35を介してディスプレイ装置42に表示する。図15に、この表示例の一例を図示する。この表示例は、ヒトチトクロームcとバクテリアチトクロームcという2つのアミノ酸配列の処理結果を表示するものであって、配列位置間の持つ文字列長分に相当する空白を挿入していくことでアラインメント処理を実行していくことになる。

【0049】

【発明の効果】以上説明したように、本発明によれば、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数や最長共有部分列を検出する構成を採って、この最長共有文字数や最長共有部分列に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミング手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に2つのアミノ酸配列の類似性を評価することができるのである。

【図面の簡単な説明】

* 【図1】本発明の原理構成図である。

【図2】本発明の一実施例である。

【図3】LCS検出部の実行する処理フローの一実施例である。

【図4】LCS検出部の実行する処理フローの一実施例である。

【図5】LCS検出部の実行する処理フローの一実施例である。

【図6】LCS検出部の作成する出現表の説明図である。

【図7】配列S[i]の更新処理の説明図である。

【図8】配列S[i]の更新処理の説明図である。

【図9】配列S[i]の更新処理の説明図である。

【図10】LCS検出部の作成するデータ構造の説明図である。

【図11】LCS検出部の実行する処理フローの一実施例である。

【図12】LCS検出部の実行する処理フローの一実施例である。

【図13】処理結果の表示形態の一実施例である。

【図14】処理結果の表示形態の一実施例である。

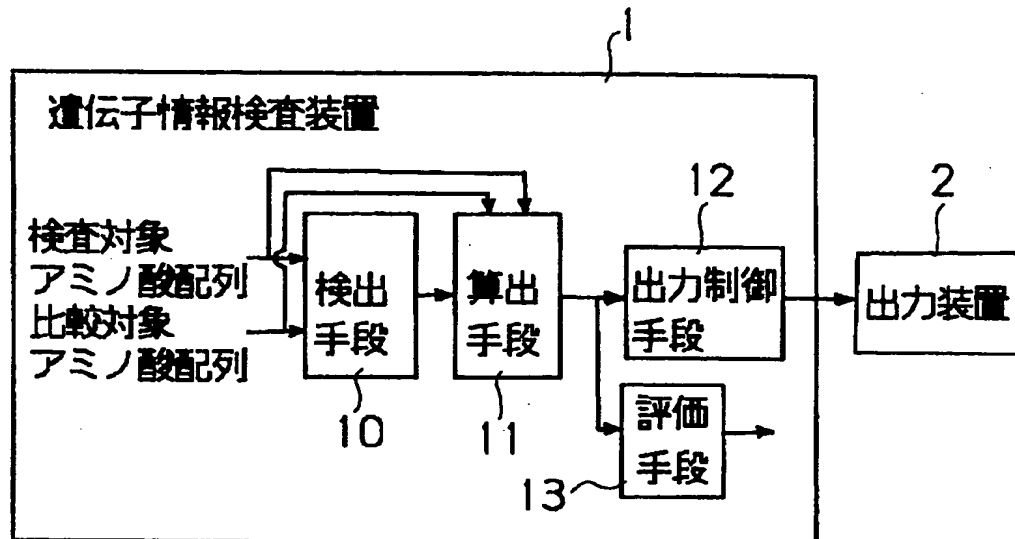
【図15】処理結果の表示形態の一実施例である。

【符号の説明】

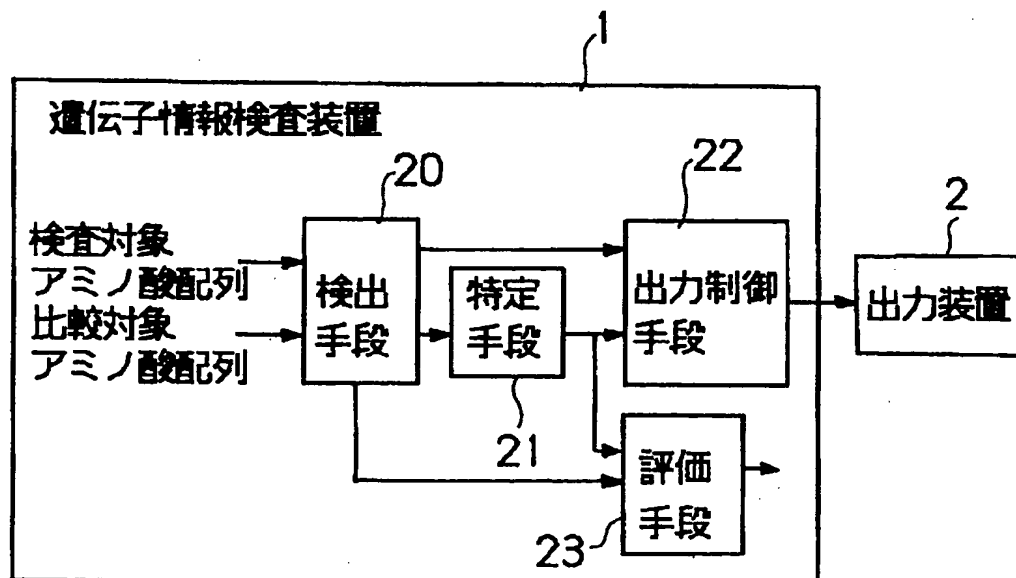
- 1 遺伝子情報検査装置
- 2 出力装置
- 10 検出手段
- 11 算出手段
- 12 出力制御手段
- 13 評価手段
- 20 検出手段
- 21 特定手段
- 22 出力制御手段
- 23 評価手段

【図1】

本発明の原理構成図



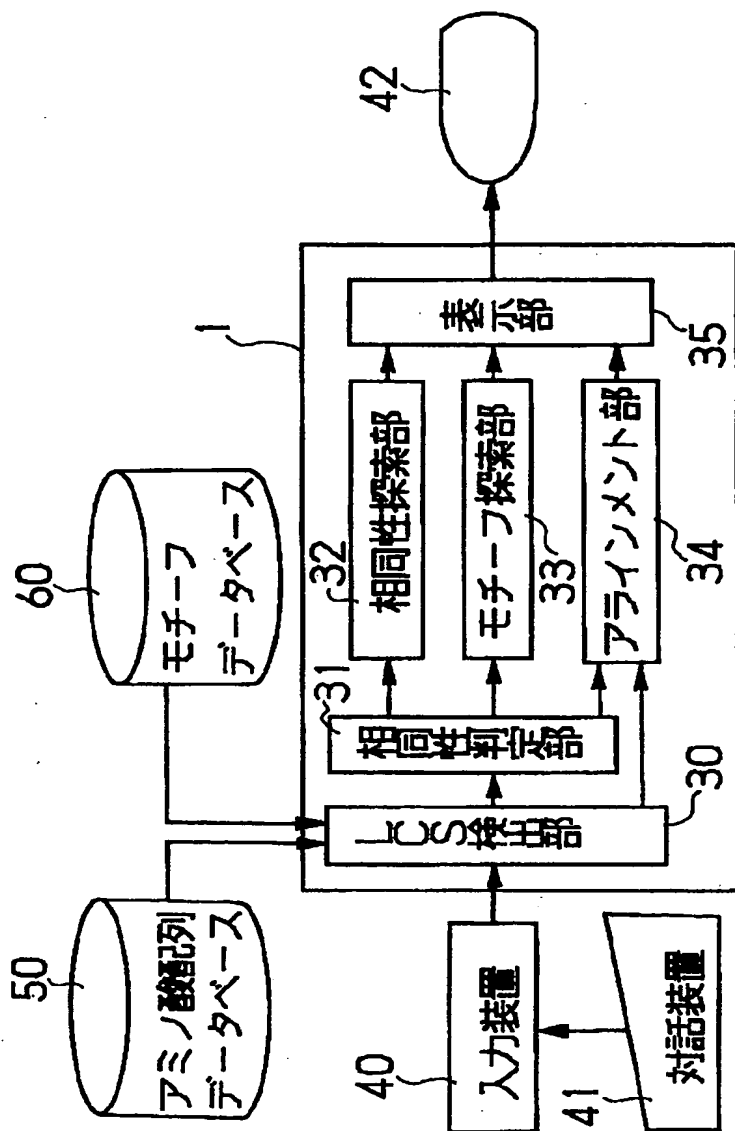
(a)



(b)

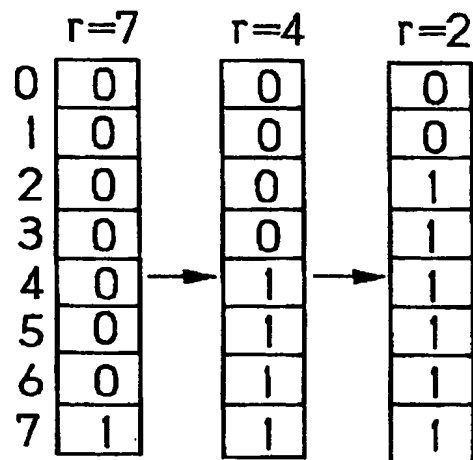
【図2】

本発明の一実施例

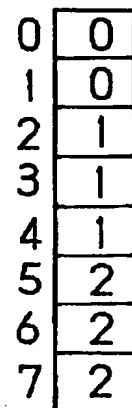


【図7】

配列S[i]の更新処理の説明図



(a)

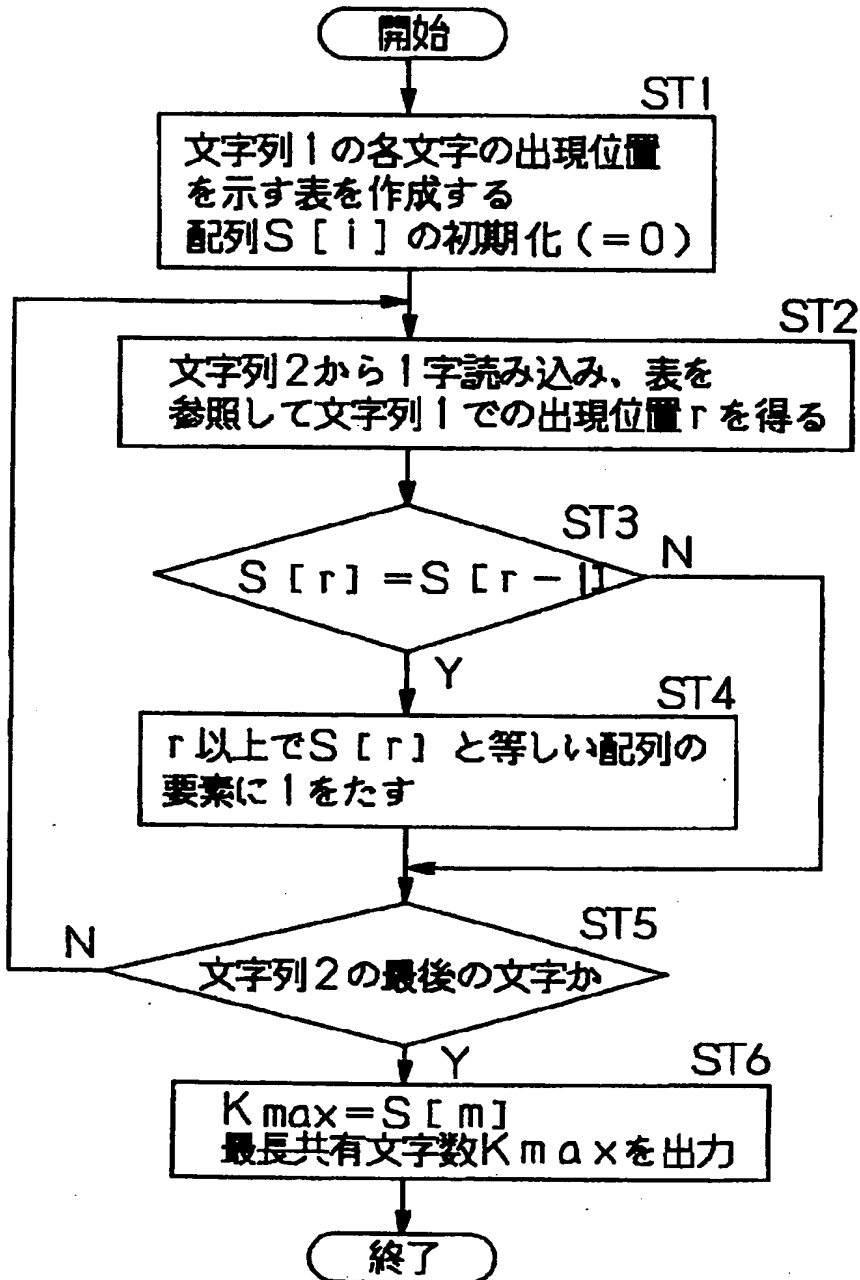


(b)

【図3】

【図8】

LCS検出部の実行する処理フローの一実施例 配列S[i]の更新処理の説明図



0	0
1	0
2	1
3	2
4	2
5	2
6	2
7	2

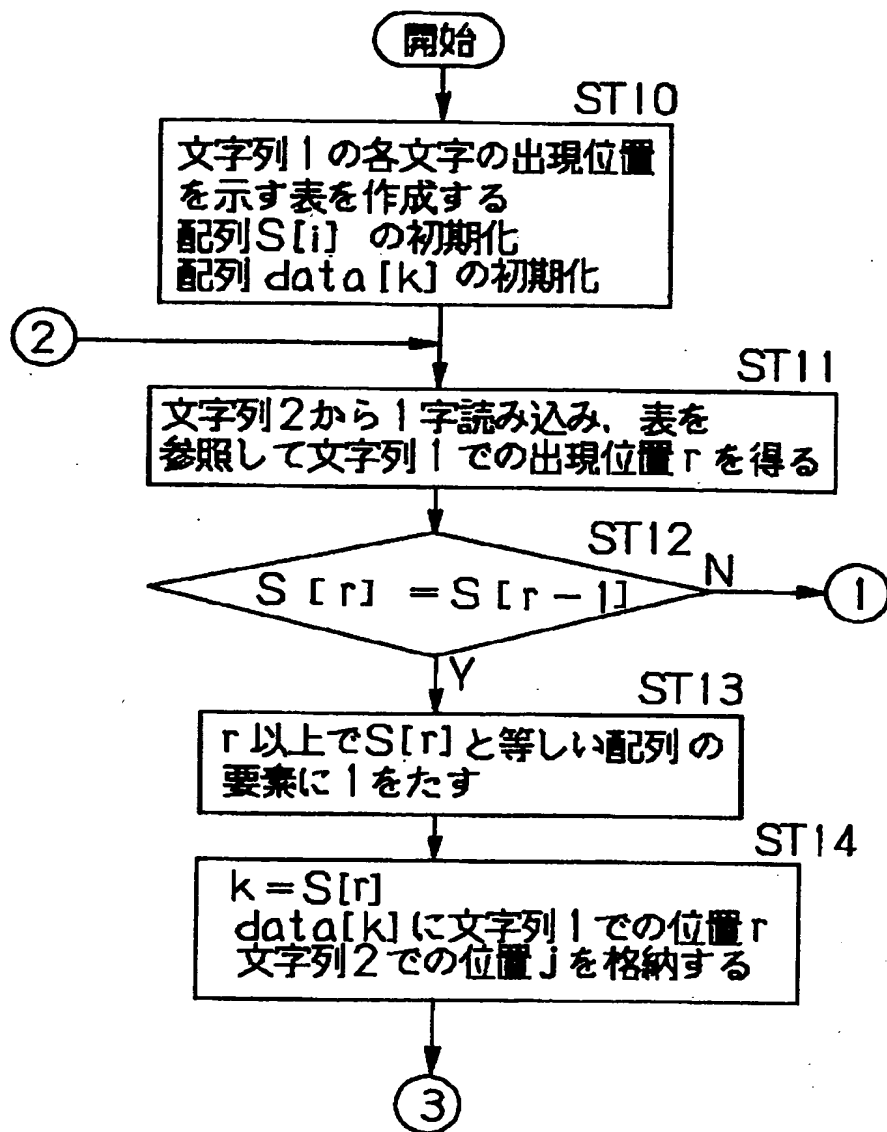
(a)

	r=6	r=1
0	0	0
1	0	1
2	1	1
3	2	2
4	2	2
5	2	2
6	3	3
7	3	3

(b)

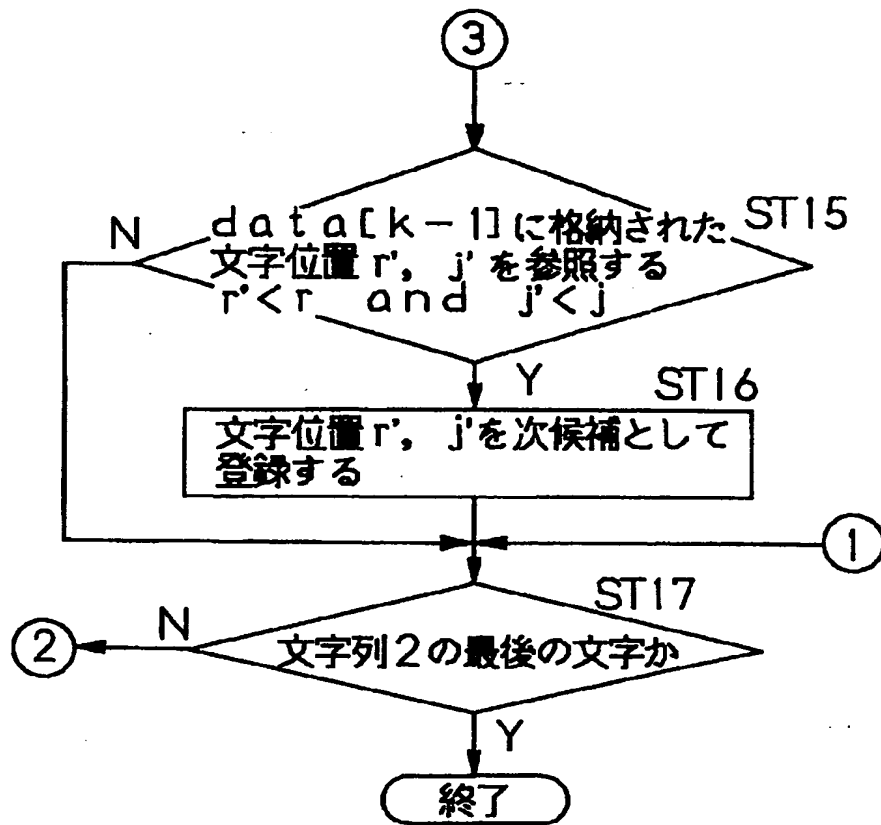
【図4】

LCS検出部の実行する処理フローの一実施例



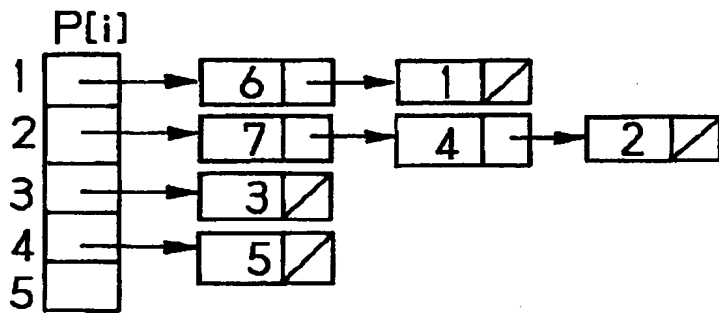
【図5】

LCS検出部の実行する処理フローの一実施例



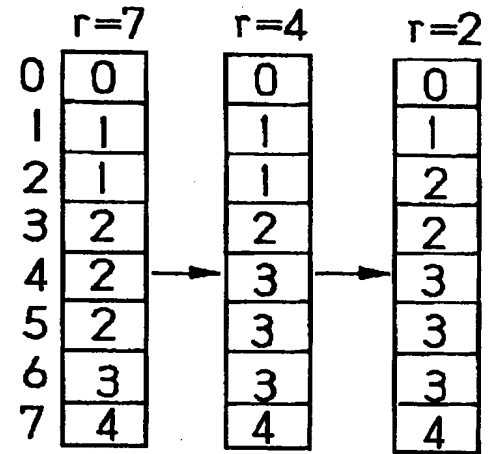
【図6】

LCS検出部の作成する出現表の説明図

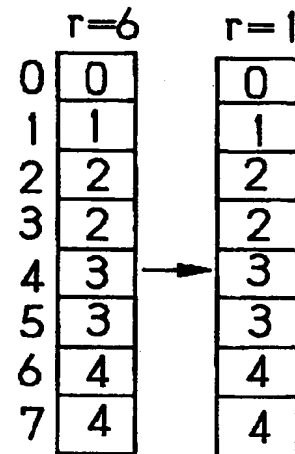


【図9】

配列S[i]の更新処理の説明図



(a)

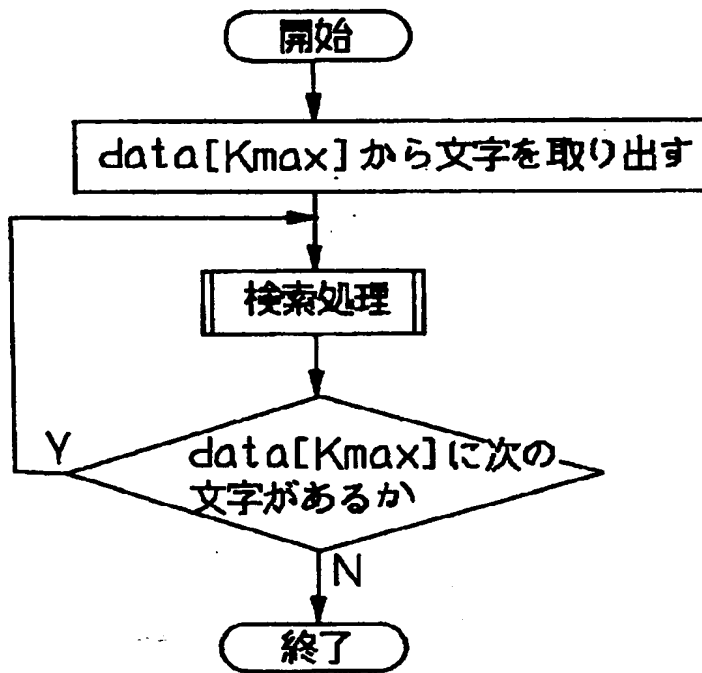


(b)

【図11】

【図15】

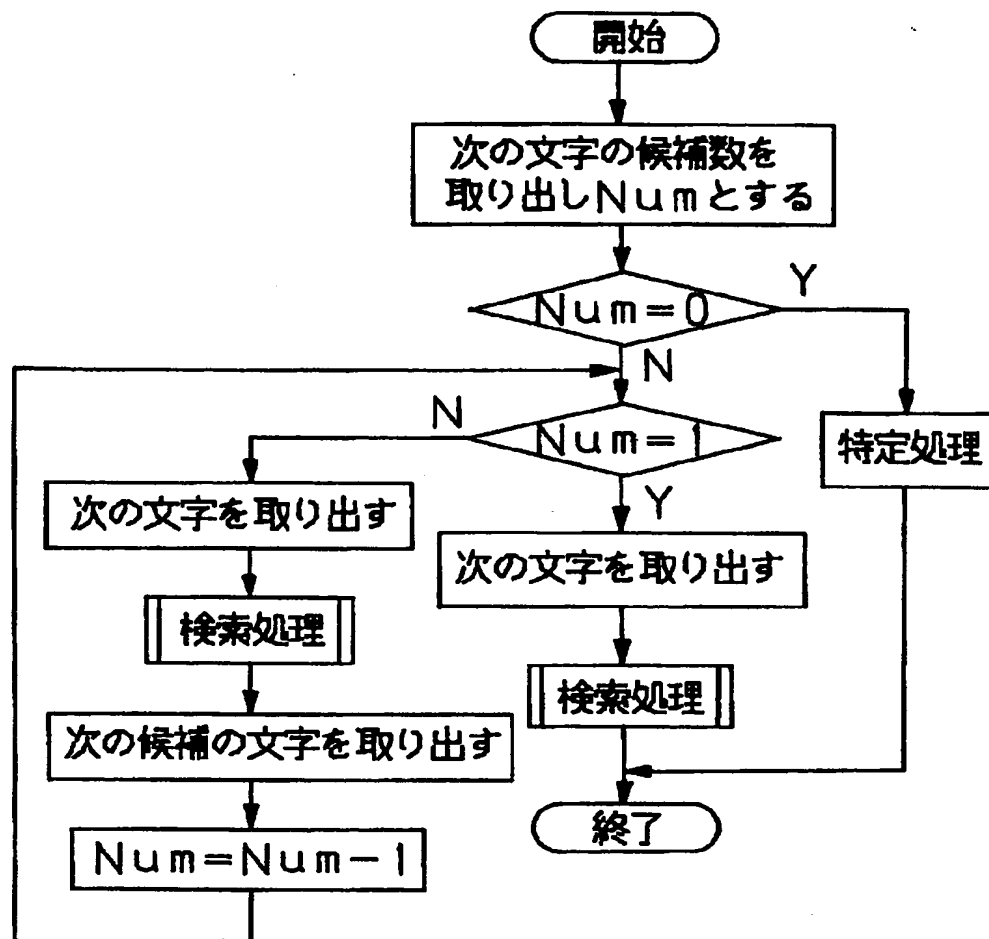
LCS検出部の実行する処理フローの一実施例 処理結果の表示形態の一実施例



human : G D V E K G K K I F I M K S Q C H T V E K G G K H K T G P N L H G L F G R K ...
bacterium: E G D A A G E K V S K K C L A C H T F D Q G G A N K V G P N P N L F G V F ...

【図12】

LCS検出部の実行する処理フローの一実施例



【図13】

処理結果の表示形態の一実施例

human : GDVEKGKKIFIMKCSQCHTVKGGKHKTGPNLHGLFGRK...

baotarium: EGDAAAGEKVSCKGLACHTFDGGANKVGPNNLFGVF...

LCS : GD(x33)G(x0,1)K(x02)K(x04)KC(x22)CHT(x33)GG(x22)K...

GD(x14)E(x02)K(x02)K(x04)KC(x22)CHT(x33)GG(x22)K...

homology: 47%